兰州大学

本科毕业论文

论文题目（中文） 基于无模型强化学习的非线性系统线性二次控制

论文题目（英文） Linear quadratic control of nonlinear systems based on model-free reinforcement learning

学生姓名 冯敏

指导教师 阎石

学　　院 信息科学与工程学院

专　　业 通信工程

年　　级 2019 级

兰州大学教务处

诚信责任书

本人郑重声明：本人所呈交的毕业论文（设计），是在导师的指导下独立进行研究所取得的成果。毕业论文（设计）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或在网上发表的论文。

本声明的法律责任由本人承担。

论文作者签名：冯敏　　日　　期：2023.5.25

关于毕业论文（设计）使用授权的声明

本人在导师指导下所完成的论文及相关的职务作品，知识产权归属兰州大学。本人完全了解兰州大学有关保存、使用毕业论文（设计）的规定，同意学校保存或向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅；本人授权兰州大学可以将本毕业论文（设计）的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本毕业论文（设计）。本人离校后发表、使用毕业论文（设计）或与该毕业论文（设计）直接相关的学术论文或成果时，第一署名单位仍然为兰州大学。

本毕业论文（设计）研究内容：

☑可以公开

□不宜公开，已在学位办公室办理保密申请，解密后适用本授权书。

（请在以上选项内选择其中一项打“√”）

论文作者签名：冯敏　　导师签名：阎石

日　　期：2023.5.25　　日　　期：2023.5.25

基于无模型强化学习的非线性系统线性二次控制

中文摘要

非线性系统的分析与控制一直以来都是控制领域的研究热点。实际控制系统通常为复杂多变的非线性系统，难以对其建立精确的数学模型，因此如何采用无模型的方法实现对非线性系统的最优控制是一个极具价值的研究方向。本文首先基于无模型强化学习理论及线性二次控制理论，采用一种改进的平均 Q-learning 算法实现了对干扰环境下线性系统的二次型最优控制。然后针对非线性系统引入 Koopman 算子理论，利用 Koopman 可观测函数将非线性系统状态向量进行升维，在升维后的线性空间中利用改进平均 Q-learning 算法迭代学习最优控制策略增益，最终在无需对系统进行建模的前提下实现了干扰环境下非线性系统的线性二次型最优控制，并通过在阻尼杜芬振荡器等非线性系统上的仿真实验验证了该方法的可行性。

关键词：非线性系统，无模型强化学习，线性二次型最优控制，Koopman 算子理论

Linear quadratic control of nonlinear systems based on model-free reinforcement learning

Abstract

The analysis and control of nonlinear systems has always been a research hotspot in the field of control. The actual control systems are usually complex and variable nonlinear systems, and it is difficult to establish an accurate mathematical model. Therefore, how to use the model-free method to achieve the optimal control of nonlinear systems is a very valuable research direction. Based on model-free reinforcement learning theory and linear quadratic control theory, this paper first uses an improved average Q-learning algorithm to achieve quadratic optimal control of a linear system in a disturbing environment. Then, the Koopman operator theory was introduced for the nonlinear system, and the Koopman observable function was used to increase the dimension of the state vector of the nonlinear system, and the improved average Q-learning algorithm was used to iteratively learn the optimal control strategy gain in the linear space after the dimension was increased. Finally, the linear quadratic optimal control of nonlinear system in disturbance environment is realized without modeling the system. The feasibility of the method is verified by simulation experiments on nonlinear systems such as damped Duffing oscillators.

Key Words: nonlinear system, Model-free reinforcement Learning, Linear quadratic optimal control,Koopman operator theory.

目　录

第一章 绪 论

1.1 研究背景及意义

非线性系统的分析和控制一直以来都是控制领域的研究要点。传统的控制方法如 PID、LQR 等在控制工程中取得了广泛的应用，在各类线性系统的控制中表现出优良的性能，但在面对非线性、强耦合以及强不确定性的被控对象时，其性能往往大幅下降。实际控制系统通常为复杂多变的非线性系统，难以建立其精确的数学模型，即使已经被理论证明具有强稳定性、收敛性和鲁棒性的方法在实际系统的控制中也可能出现诸多问题，因此无模型控制方法引起了广泛关注并快速发展起来 [1]。

近年来，强化学习（RL）在许多具有挑战性的控制任务中表现出了优异的性能 [2]。其中无模型强化学习在无需对未知动力学系统进行建模的情况下即可找到最优控制策略。无模型强化学习方法是一种“端到端”的方法，其直接优化成本函数，避免了建模的困难且易于实现 [3]。因此无模型强化学习十分适合应用于对难以建立系统模型的复杂非线性系统进行控制。无模型强化学习的目标是在不对系统进行显式建模的情况下找到最优策略，其通过与环境交互获取数据估计值函数或直接优化参数化策略，虽易于实现但相比于基于模型的方法缺乏理论保障，使得在将其部署到实际系统中时存在隐患。

线性二次型（LQ）问题是一类经典的控制问题，其研究的系统为线性动力学系统且要最小化的成本函数是二次型的，具有解析解。由于线性二次型问题具有强的理论保证且广泛应用于各种工程问题，具有强的实用性，因此将无模型强化学习与线性二次型问题相结合研究动力学系统的最优控制具有良好的发展前景。Matni, Nikolai 等人 [4] 调研了针对线性二次型问题的 RL 方法，分别从基于模型和无模型两个方面进行了总结，本文主要关注基于无模型 RL 方法的线性二次调节（LQR）控制问题。例如，Bian, Tao 等人 [5] 考虑了带有过程噪声的线性二次型问题，假设噪声是可测量的，并提出了一种最小二乘时间差分学习（LSTD）算法来实现最优控制。Bahare 等人 [6] 采用一种无模型异策略强化学习方法给出了线性离散时间系统的 H_∞ 控制的解。Abbasi-Yadkori 等人 [3] 针对带过程噪声的线性二次型问题提出了改进的 Q-learning 算法，并分析了算法的遗憾界。Yaghmaie 等人 [2] 考虑了带有未知的随机过程噪声和测量噪声的 LQR 问题，并提出了两种改进的策略迭代算法学习最优控制策略。上述工作均是基于线性动力学系统进行研究，所提出的无模型 RL 算法可针对线性动力学系统实现最优反馈控制，本文主要研究非线性动力学系统的最优控制问题，因此需要将非线性系统进行全局线性化，然后利用解决 LQR 问题的无模型强化学习算法实现最优反馈控制。

1931 年提出的 Koopman 算子理论近年来成为非线性系统的线性表示的重要方法。Koop-

man 算子可以将非线性系统映射到无穷维线性空间上，通过无穷维线性空间的有限维近似实现对非线性系统的全局线性化[7]。利用 Koopman 可观测函数可以将非线性动力学系统的动态演化表示为一个无穷维的线性变换，从而可以使用线性系统理论来分析非线性系统的动态行为[8]。通过在升维后的空间中设计控制器可以实现对原非线性动力学系统的精确控制。因此本文基于 Koopman 算子理论对非线性系统进行升维，利用 Koopman 可观测函数将非线性系统的观测数据映射到一个高维空间中，在这个升维后的空间中采用解决 LQR 问题的无模型强化学习算法学习最优控制策略，从而实现对原非线性系统的最优控制。本文提出的基于无模型强化学习的非线性系统线性二次控制方法在无需对系统进行建模的情况下实现了非线性系统的最优反馈控制，与一般的无模型方法不同，其基于 LQR 问题构建值函数及贝尔曼方程，具有更强的理论保证，可以很好地解决带有未知干扰的非线性动力学系统的最优控制问题。

1.2 国内外研究现状

1.2.1 非线性系统研究现状

非线性系统由于具有混沌现象、周期运动等复杂的动态行为使得其控制变得十分困难，因此，非线性系统的控制研究一直是控制领域的热点和难点问题。常用的非线性系统控制方法有 PID 控制方法、自适应控制方法、滑模控制方法以及近年来得到快速发展的基于机器学习的控制方法等。Iqbal, J. 等人[9] 总结了非线性控制系统的发展与挑战，并从应用角度考察了非线性控制系统的进展。Mukhtar Fatihu Hamza 等人[10] 在旋转倒立摆上测试比较了线性、非线性时不变、自学习和自适应非线性控制器的性能。

其中自适应动态规划（ADP）是处理最优控制最有效的方法之一，利用一个函数近似结构来估计代价函数并求解动态方程获得近似最优控制策略[11]。Werbos 等人[12] 提出 ADP 的概念用以解决“维度灾难”的问题，利用函数近似结构来估计动态方程中出现的性能指标和控制策略，取得了良好的控制效果。近年来，随着人工智能的蓬勃发展，许多学者将经典的最优控制理论与先进的人工智能技术相结合，取得了诸多成果。例如，Wen[13] 等人提出了一种基于自适应神经网络强化学习的跟踪优化控制方法，构造 actor-critic 网络来估计性能指标函数。闫珍珍[14] 基于 BP 神经网络提出了两种受扰非线性系统的控制方法。Ouyang 等人[15] 研究了给定约束条件下机械臂 (RMs) 的最优控制问题，将误差转换技术与 ADP 融合提出近似最优控制策略，并在二自由度机械臂上进行仿真实验测试了控制器的性能。范昕[11] 按照从单个系统到多智能体系统、单输入到多输入和一般到特殊的逻辑，分别考虑单输入单输出系统、多输入系统、严格反馈系统和多智能体系统提出 ADP 控制策略。

1.2.2 强化学习研究现状

强化学习是一种通过智能体与环境交互试错从而学习如何做出正确决策以最大化累积奖励的机器学习方法。强化学习的研究始于 20 世纪 50 年代，1998 年 Sutton 等人[16] 在书中系统地介绍了强化学习的理论和算法，创作了强化学习领域的经典著作。强化学习主要可以分为基于模型和无模型两大类别，其中基于模型的方法需要事先知道完整的环境模型，而无模型方法则是直接通过交互数据学习控制策略。孙悦雯等人[17] 研究了基于因果建模的强化学习控制，并对其研究现状进行了总结和展望，本文主要关注无模型强化学习算法。

无模型强化学习算法中包含基于价值的强化学习算法和基于策略的强化学习算法。其中前者通过估计价值函数及状态动作值函数进而更新策略，Watkins[18] 提出的 Q-learning 算法是一个经典的基于价值的强化学习算法，其通过学习状态价值表选取能获得最大收益的动作。为解决状态价值表难以处理高维状态空间的问题，DeepMind 团队提出了 DQN 算法[19]，该算法首次实现了用神经网络估计值函数，促进了深度强化学习（DRL）的发展，后续在该算法的基础上提出了诸多改进算法，如 Double DQN、Rainbow 等。策略梯度（PG）算法是基于策略的经典强化学习算法，其直接学习策略输出动作的概率，可以解决连续动作空间问题。因此将基于价值的算法与策略梯度思想相结合提出了 actor-critic 架构，既学习价值函数又学习策略函数，为后续诸多常用强化学习算法奠定了基础。例如 Schulman, J 等人[20] 提出近端策略优化（PPO）算法，采用 AC 框架，用高斯分布对连续动作空间建模并通过采样获得输出动作，获得了广泛的应用。Lillicrap 等人[21] 提出的深度确定性策略梯度（DDPG）算法采用神经网络拟合值函数和策略函数，可以有效处理高维状态及动作空间问题。Haarnoja, T 等人[22] 提出的 SAC 算法引入了最大熵的概念，增强了算法的探索能力，提升了算法的鲁棒性及泛化能力，在诸多应用中取得了良好的表现。

随着无模型强化学习算法的不断发展与改进，其在控制领域取得了广泛的应用，具有十分良好的发展前景。Gu 等人[23] 证明了一种基于深度 Q 函数的离线深度强化学习算法可以可以扩展到复杂的 3D 操作任务，可以有效地学习深度神经网络策略，可以在真实的物理机器人上进行训练，并通过实验证明该方法可以让实际机器人学习复杂的开门技能。杨永亮[24] 针对领航者带有未知控制输入的异构多智能体系统输出同步控制问题，提出了无模型自适应动态规划，并针对具有多个领航者的异构多智能体系统的包含控制问题，利用无模型自适应动态规划，设计了完全分布式的最优包含控制律。Xi 等人[25] 将基于模型的强化学习 (MBRL) 和无模型强化学习 (MFRL) 相结合，提出了一种基于 $Q(\lambda)$ 的 MFRL 方法改进策略，仿真结果表明所提 RL 算法能够实现 NAO 机器人在旋转平台上的平衡，并能适应平台角速度的变化。许雅筑等人[26] 介绍了强化学习在镇定控制和跟踪控制等底层控制任务方面的应用，重点对强化学习在自主水下机器人控制领域面临的挑战以及最新的进展进行了概述，并详细介绍了两种针对自主水下机器人的无模型强化学习控制方法。

1.2.3 Koopman 算子理论研究现状

Koopman 算子的研究始于 20 世纪 30 年代，由数学家 Koopman, B. O. 首先提出 Koopman 算子理论 [27]。通过 Koopman 算子可以将动力学系统中的非线性演化转变为线性演化，从而可以使用线性系统相关理论进行数学分析和控制设计，因此 Koopman 算子广泛应用于非线性系统控制领域中。Budii, Marko 等人 [28] 梳理总结了 Koopman 算子的相关理论，以简洁的方式描述 Koopman 理论的框架以便于后续研究利用，并分析了其理论研究到应用实践的潜力。

Rowley, C. W. 等人 [29] 利用基于数据驱动的动态模式分解方法来寻找 Koopman 的特征值和模态，并以横流中的射流为例说明了该方法，表明该方法捕获了主导频率并阐明了相关的空间结构。Williams 等人 [30] 进一步研究了 Koopman 模态与 DMD 方法，提出了一种扩展动态模态分解（EDMD）方法用来计算 Koopman 算子的特征值、本征函数及模态，通过示例突出了该方法在确定性或随机性数据下的性能，并展示了 Koopman 本征函数的应用潜能。Korda 等人 [8] 提出了一种新的数据驱动框架，该方法利用 Koopman 算子远离吸引子的频谱的丰富性来构造一组本征函数，利用优化构造的方法从数据中学习面向预测和控制的 Koopman 本征函数并构建线性预测器，将所构建的线性预测器与模型预测控制（MPC）方法相结合，在阻尼杜芬振荡器等非线性系统上测试了该方法的性能。随着深度学习等技术的发展，利用深度学习来解决传统的 Koopman 算子研究中维度灾难等问题的方法引起了学者们的注意，Bethany 等人 [31] 利用深度学习算法从数据中寻找 Koopman 本征函数，提出结构简洁且可解释的网络将动力学嵌入到低维流形上，用改进的自编码器来识别其上动力学是全局线性的非线性坐标。Koopman 算子研究的不断发展与完善使其应用范围也不断拓展，为人们更好地理解和控制复杂系统提供了极大的帮助。

1.3 本文主要工作

本文主要研究如何基于无模型强化学习实现带有未知干扰的非线性动力学系统的线性二次控制。本文采用了改进的 Q-learning 算法解决线性系统的最优反馈控制问题，并进一步基于 Koopman 算子理论对非线性系统进行全局线性化，然后利用实现的无模型强化学习线性二次控制算法对非线性系统进行最优控制。本文的章节安排如下：

第一章为绪论。首先分析了研究基于无模型强化学习的非线性系统线性二次控制方法的背景及意义。然后总结了非线性系统控制、强化学习以及 Koopman 算子理论的研究现状。最后简要介绍了本文的主要工作及安排。

第二章为理论及算法基础。首先介绍了强化学习理论，对马尔可夫决策过程和无模型强化学习的基本概念进行了阐述。然后介绍了经典的线性二次控制理论，最后介绍了 Koopman 算子理论以及受控系统下 Koopman 算子的推广。

第三章为基于 Q-learning 的线性二次控制。首先介绍了线性二次型最优控制问题，然

后基于该问题的基本概念定义了强化学习中的价值函数及状态-动作值函数。然后着重描述了一种改进的平均 Q-learning 算法实现线性系统的线性二次最优控制。最后列举了数值例子测试该算法的性能。

第四章为基于 Koopman-无模型强化学习的非线性系统线性二次控制。本章基于 Koopman 算子理论及第三章中的改进平均 Q-learning 算法实现非线性系统的线性二次最优控制。首先介绍了基于 Koopman 理论的非线性系统状态升维方法，其中重点描述了如何利用最优化方法构造 Koopman 本征函数。然后用将 Koopman 理论引入第三章中的无模型强化学习算法实现对非线性系统的控制。最后在数值例子上测试了所提出方法的性能。

第五章为总结与展望。本章对本文的研究内容及结果进行了总结并对之后的研究内容进行了展望。

第二章 理论及算法基础

本章简要介绍了后续章节使用到的理论知识及算法基础。首先介绍了强化学习理论，对马尔可夫决策过程和无模型强化学习进行了简单描述；然后引入了经典的线性二次控制理论，简要介绍了线性高斯动力系统和线性二次型最优控制问题的基本概念，并对 Riccati 方程进行了详细的推导；最后针对非线性系统最优控制问题，引入了 Koopman 算子理论，介绍了 Koopman 算子及可观测函数。

2.1 强化学习理论

强化学习讨论的问题是一个智能体(agent)如何在一个复杂不确定的环境(environment)中去极大化它能获得的奖励[32]。随着人工智能的兴起，强化学习的发展突飞猛进，融入深度学习的深度强化学习更是在自适应优化控制、机器人控制等诸多领域都取得了可观的成果。

强化学习通过智能体不断与环境进行交互，利用交互得到的数据学习可以获得最大的长期奖励的状态到动作的映射。强化学习有三个基本的元素，即状态、动作和奖励。强化学习的基本模型如图2.1所示，智能体采取动作与环境进行交互，从环境获得状态和相应的奖励数据，利用得到的数据学习价值函数或策略函数并进一步得到使长期奖励最大化的控制策略。强化学习依据状态转移概率或状态转移函数是否已知分为有模型和无模型两大类别，其中无模型强化学习在无需对未知动力学系统进行建模的情况下即可找到最优控制策略，避免了对复杂不确定系统的建模问题。

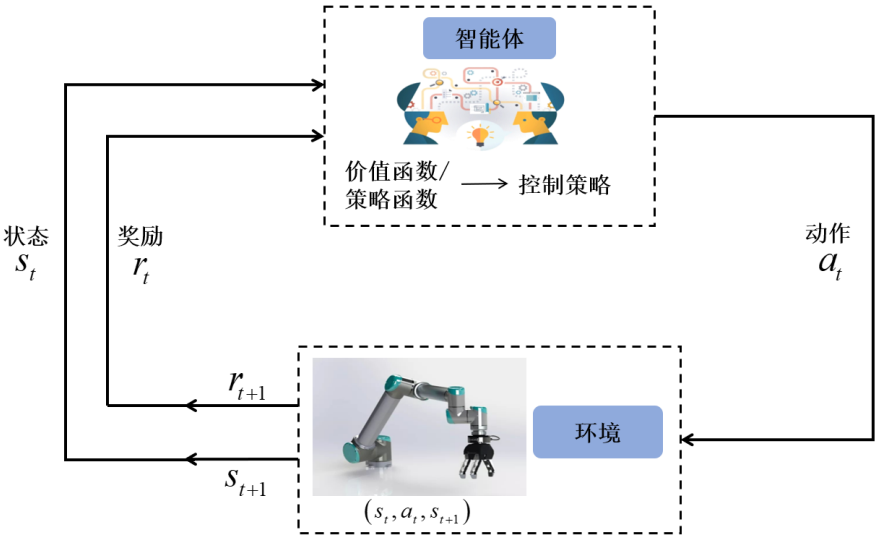

图 2.1 强化学习基本模型

2.1.1 马尔可夫决策过程

马尔可夫决策过程（MDP）是强化学习中一个重要的基本概念，利用强化学习解决实际问题前首先要使用马尔可夫决策过程来表述实际问题。马尔可夫决策过程定义为具有马尔可夫转移模型和附加奖励的完全可观察的随机环境的顺序决策问题[33]。其中马尔可夫转移模型表明该问题某一时刻的状态只与上一时刻的状态有关，与之前的状态无关，可以用状态转移函数 $P(s'|s,a)$ 表示。

MDP 通过数学方式表述了强化学习中智能体与环境交互的过程，由五元组 $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ 构成。其中状态空间 $\mathcal{S}$ 是所有可能的状态的集合；动作空间 $\mathcal{A}$ 是智能体所有可能采取的动作的集合；状态转移函数数 $P(s'|s,a)$ 表示在状态 s 下采取动作 a 转移到状态 s' 的概率；奖励函数 $r(s,a)$ 是在状态 s 下采取动作 a 获得的奖励；折扣因子 γ 在计算累积回报时用于表示对远期奖励的关注程度，其取值范围为 $[0,1]$。将智能体的策略 $\pi(a|s)$ 定义为在状态 s 下智能体采取动作 a 的概率，确定性策略在每个状态下只采取一个确定的动作，即该动作的概率为 1，本文采用的就是确定性策略。强化学习的目标就是对于给定的马尔可夫决策过程寻找最优即最大化累积奖励期望的策略。

给定策略 π，t 时刻状态 s_t 下的累积奖励定义为：

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \tag{2.1}$$

其中，R_t 表示在时刻 t 获得的奖励。一个状态的价值定义为该状态的期望回报，所有状态的价值就组成了价值函数[34]：

$$\begin{aligned} V^{\pi}(s) &= \mathrm{E}_{\pi}[G_t | s_t = s] \\ &= \mathrm{E}_{\pi}\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots | s_t = s\right] \\ &= \mathrm{E}_{\pi}\left[R_t + \gamma\left(R_{t+1} + \gamma R_{t+2} + \cdots\right) | s_t = s\right] \\ &= \mathrm{E}_{\pi}\left[R_t + \gamma G_{t+1} | s_t = s\right] \\ &= \mathrm{E}_{\pi}\left[R_t + \gamma V^{\pi}(s_{t+1}) | s_t = s\right] \end{aligned} \tag{2.2}$$

价值函数的输入为状态，输出为该状态对应的价值，式2.2为价值函数的贝尔曼方程。

状态动作值函数表示 MDP 在状态 s 下根据策略 π 执行动作 a 得到的期望回报，其数学表达如下：

$$\begin{aligned} Q^{\pi}(s,a) &= \mathrm{E}_{\pi}[G_t | s_t = s, a_t = a] \\ &= \mathrm{E}_{\pi}\left[R_t + \gamma Q^{\pi}(s_{t+1}, a_{t+1}) | s_t = s, a_t = a\right] \end{aligned} \tag{2.3}$$

式2.3是状态动作值函数的贝尔曼方程。

2.1.2 无模型强化学习

对于马尔可夫决策过程问题，在模型已知的情况下，可以利用动态规划来评估给定的策略并通过不断迭代得到最优价值函数，但大部分应用强化学习的现实场景如机器人控制等，往往具有复杂的物理环境，其马尔可夫决策过程中的状态转移概率或状态转移函数难以获得，即难以对系统进行准确建模，无法直接利用贝尔曼方程求解得到最优策略。因此无模型强化学习的重要性不言而喻，其在无需对系统动力学进行建模的情况下通过与环境不断交互获取到的状态及奖励数据进行学习，这使得其可以被广泛应用于实际场景中。

时序差分（TD）是无模型强化学习中的一种重要方法，用来估计一个策略的价值函数，它结合了蒙特卡洛和动态规划的思想[34]。时序差分方法在事先不知道环境的情况下从样本数据中进行学习，根据贝尔曼方程的思想，利用后续状态的价值估计更新当前状态的价值估计。对于模型已知的情况，可直接进行动态规划，值函数通过下式计算：

$$V(s_t) \leftarrow E_\pi\left[R_{t+1}+\gamma V(s_{t+1})\right] = \sum_a \pi(a \mid s_t) \sum_{s',r} p(s',r \mid s_t,a)\left[r+\gamma V(s')\right]$$

上式中使用了当前状态 s_t 的所有后继状态 s_{t+1} 处的值函数来估计 s_t 处的值函数，即采用了自举的方法。而当无模型时，无法知道所有的后继状态，此时可通过试验采样的方法利用多次试验的经验平均值来估计值函数，蒙特卡洛方法就是采用该思想，其值函数通过下式进行估计：

$$V(s_t) \leftarrow V(s_t)+\alpha\left(G_t - V(s_t)\right)$$

上式中 G_t 表示从状态 s_t 开始完成一次试验后到终止状态的所有回报的值，即蒙特卡洛方法需要等完成一次完整试验才能更新值函数。

时序差分方法结合了动态规划的自举以及蒙特卡洛方法的采样方法，其值函数更新方式如式2.4所示。

$$V(s_t) \leftarrow V(s_t)+\alpha\left(R_{t+1}+\gamma V(s_{t+1}) - V(s_t)\right) \tag{2.4}$$

时间差分方法用当前获得的实际奖励与其后继状态的价值估计相加作为当前状态的回报，α 表示对价值估计更新的步长，可取为一个常数。时序差分目标 $R_{t+1}+\gamma V(s_{t+1})$ 是带衰减的长期奖励的总和，利用自举的方法来估计 $V(s_{t+1})$，$R_{t+1}+\gamma V(s_{t+1})-V(s_t)$ 被称为时序差分误差（TD error），其与更新步长的乘积作为时序差分算法的更新量。

Q-learning 是一种常用的经典基于价值迭代的无模型强化学习算法，其采用 TD 算法来估计动作状态价值函数 Q，更新方式如式 2.5所示。

$$Q(s_t,a_t) = Q(s_t,a_t)+\alpha\left[R_{t+1}+\gamma \max Q(s_{t+1},a) - Q(s_t,a_t)\right] \tag{2.5}$$

Q-learning 是离线策略（off-policy）算法，为了探索，其通常采用 ε-贪婪策略与环境交互获取数据用于更新，且用于交互的行为策略与用于更新的目标策略不是同一个策略，如此使得算法具有更小的样本复杂度，更加被广泛使用。

2.2 线性二次控制理论

线性二次型最优控制是经典的控制问题，其中动力学系统服从线性动力学，并且要最小化的代价函数是二次型的，具有强的理论保证和可实现性，在实践中得到了广泛应用。由于无模型强化学习与基于模型的方法相比理论保证更少，因此将线性二次控制理论与无模型强化学习相结合，为其提供强的理论保证，实现更稳定高效的无模型控制算法。

2.2.1 线性二次型最优控制问题

线性二次型最优控制的目的是在一定的性能指标下，使系统的控制效果最佳，利用最少的控制能量，来达到最小的状态误差[35]。线性离散系统的状态方程和成本函数定义如下：

$$x_{k+1} = Ax_k + Bu_k \tag{2.6}$$

$$J = \sum_{k=0}^{\infty} \left({x_k}^{\mathrm{T}} R_x x_k + {u_k}^{\mathrm{T}} R_u u_k \right) \tag{2.7}$$

其中 x_k 和 u_k 分别表示状态和控制输入向量，R_x 和 R_u 分别是状态和控制加权矩阵，通过选取不同 R_x、R_u 加权矩阵得出的指标函数 J 表示输出量衰减速度与控制过程消耗能量之间的折中[35]。R_x 的值越大，表明对其对应的状态误差越重视，该状态衰减越快，但同时消耗能量越大，R_u 的值越大表明牺牲状态衰减速度节约能量输入。设计状态反馈控制器 $u_k = -Kx_k$，最优控制器使成本函数最小。

2.2.2 离散 LQR 最优控制问题求解

可利用拉格朗日法求解线性二次型最优控制问题，将成本函数变化为如下形式：

$$J = \sum_{k=0}^{\infty} \left({x_k}^{\mathrm{T}} R_x x_k + {u_k}^{\mathrm{T}} R_u u_k \right) + {x_n}^{\mathrm{T}} R_x x_n$$

引入拉格朗日乘子 λ，得到

$$\begin{aligned} J &= \sum_{k=0}^{n-1} \left({x_k}^{\mathrm{T}} R_x x_k + {u_k}^{\mathrm{T}} R_u u_k \right) + {x_n}^{\mathrm{T}} R_x x_n + \sum_{k=0}^{n-1} \lambda_{k+1}^{\mathrm{T}} \left(Ax_k + Bu_k - x_{k+1} \right) \\ &= \sum_{k=0}^{n-1} \left({x_k}^{\mathrm{T}} R_x x_k + {u_k}^{\mathrm{T}} R_u u_k + \lambda_{k+1}^{\mathrm{T}} \left(Ax_k + Bu_k \right) - \lambda_{k+1}^{\mathrm{T}} x_{k+1} \right) + {x_n}^{\mathrm{T}} R_x x_n \end{aligned}$$

令 $H_k = x_k^T R_x x_k + u_k^T R_u u_k + \lambda_{k+1}^T \left(Ax_k + Bu_k \right)$，可以得到

$$\begin{aligned} J &= \sum_{k=0}^{n-1} \left(H_k - \lambda_{k+1}^T x_{k+1} \right) + x_n^T R_x x_n \\ &= \sum_{k=0}^{n-1} \left(H_k - \lambda_k^T x_k \right) + x_n^T R_x x_n - \lambda_n^T x_n + \lambda_0^T x_0 \end{aligned}$$

对上式的成本函数进行求导运算，可得到以下式子：

$$\frac{\partial J}{\partial x_k}=\vec{0} \quad \Rightarrow \quad \lambda_k=\frac{\partial H_k}{\partial x_k} \tag{2.8a}$$

$$\frac{\partial J}{\partial u_k}=\vec{0} \quad \Rightarrow \quad \frac{\partial H_k}{\partial u_k}=\vec{0} \tag{2.8b}$$

$$\frac{\partial J}{\partial x_n}=0 \quad \Rightarrow \quad \frac{\partial\left(x_n^{\mathrm{T}}R_x x_n-\lambda_n^T x_n\right)}{\partial x_n}=0 \tag{2.8c}$$

$$\frac{\partial J}{\partial \lambda_{k+1}}=\vec{0} \quad \Rightarrow \quad Ax_k+Bu_k-x_{k+1}=\vec{0} \tag{2.8d}$$

进一步求解可以得到：

$$\lambda_k=2R_x x_k+A^T\lambda_{k+1} \tag{2.9a}$$

$$u_k=-\frac{1}{2}R_u^{-1}B^T\lambda_{k+1} \tag{2.9b}$$

$$\lambda_n=2R_x x_n \tag{2.9c}$$

$$x_{k+1}=Ax_k+Bu_k \tag{2.9d}$$

将2.9c代入2.9b的 $k=n-1$ 式得到 $u_{n-1}=-\frac{1}{2}R_u{}^{-1}B^T\left(2R_x x_n\right)=-R_u{}^{-1}B^T R_x x_n$，将该式代入2.9d的 $k=n-1$ 可得到

$$x_n=Ax_{n-1}+B\left(-R_u{}^{-1}B^T R_x x_n\right)\Rightarrow x_n=\left(I+BR_u{}^{-1}B^T R_x\right)^{-1}Ax_{n-1} \tag{2.10}$$

将2.9c代入2.9a的 $k=n-1$ 式得到

$$\lambda_{n-1}=2R_x x_{n-1}+A^T\lambda_n=2R_x x_{n-1}+A^T 2R_x x_n \tag{2.11}$$

将2.10代入上式得到

$$\begin{aligned}\lambda_{n-1}&=2R_x x_{n-1}+2A^T R_x\left(I+BR_u{}^{-1}B^T R_x\right)^{-1}Ax_{n-1}\\&=2\left(R_x+A^T R_x\left(I+BR_u{}^{-1}B^T R_x\right)^{-1}A\right)x_{n-1}\end{aligned} \tag{2.12}$$

假设 $\lambda_k=2P_k x_k$，由2.9c可得 $P_n=R_x$，重新进行上述递推过程有

$$\lambda_{n-1}=2\left(R_x+A^T P_n\left(I+BR_u{}^{-1}B^T R_x\right)^{-1}A\right)x_{n-1}=2P_{n-1}x_{n-1}$$

，由此转换可得到递推式，即离散代数 Riccati 方程

$$P=Q+A^T PA-A^T PB\left(R+B^T PB\right)^{-1}B^T PA \tag{2.13}$$

计算控制输入

$$u_k=-\frac{1}{2}R_u{}^{-1}B^T\lambda_{k+1}=-R_u{}^{-1}B^T P_{k+1}\left(Ax_k+Bu_k\right)\Rightarrow u_k=-\left(R_u+B^T P_{k+1}B\right)^{-1}B^T P_{k+1}Ax_k$$

由此可以得到反馈增益：

$$K=\left(R_u+B^T P_{k+1}B\right)^{-1}B^T P_{k+1}A \tag{2.14}$$

由上述推导可知，对于状态即控制转移矩阵及 A 和 B 已知的线性系统，其线性二次型最优控制具有解析解，而当 A 和 B 未知时，则无法直接求解，此时可利用无模型强化学习方法学习最优反馈增益。

2.3 Koopman 算子理论

库普曼算子是一种线性算子，它控制标量函数 (通常被称为可观测值) 沿给定非线性动力系统轨迹的演化 [36]。Koopman 理论可以将非线性系统升维到无限维的线性空间中，即用作用于系统状态测量函数的希尔伯特空间的无限维线性算子来表示非线性动力学系统，从而使用线性系统控制方法进行控制。

2.3.1 Koopman 算子定义及可观测函数

Koopman 证明了可以用一个作用于系统状态的测量函数(通常也被称为可观测函数)的希尔伯特空间的无穷维线性算子来表示一个非线性动力学系统。该算子被命名为 Koopman 算子，它的谱分解完全表征了非线性动力学的行为。

对于离散自治动力学系统：

$$x_{k+1} = F(x_k)$$

Koopman 算子的定义如下：

$$\mathcal{K}g(x_k) \triangleq g(F(x_k)) = g(x_{k+1}) \tag{2.15}$$

其中 g 为该系统的一组 Koopman 可观测函数。由于 Koopman 算子是无限维的线性算子，因此应用 Koopman 分析并非捕捉希尔伯特空间中所有可观测函数的演变，而是试图识别随着动力学流动线性演变的关键可观测函数。Koopman 算子的本征函数提供了这样一组在时间上表现为线性的特殊可观测函数，koopman 算子 $\mathcal{K}$ 的本征函数 ϕ 和其相应的特征值 λ 有如下关系：

$$\mathcal{K}\phi(x_k) = \lambda\phi(x_k) = \phi(x_{k+1}) \tag{2.16}$$

对于连续时间动力学系统有：

$$\frac{d}{dt}g = \mathcal{K}g \tag{2.17}$$

其 Koopman 算子的本征函数 ϕ 和其相应的特征值 λ 满足下式：

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(\mathbf{x}) = \mathcal{K}\phi(\mathbf{x}) = \lambda\phi(\mathbf{x}) \tag{2.18}$$

利用本征函数进行扩展可以得到更多的可观测函数，即将 g 表示为本征函数的线性组合则有：

$$g(x) = \sum_{i=1}^{n} a_i\phi_i(x) \tag{2.19}$$

从数据或解析表达式中获得 Koopman 本征函数是现代动力学系统中的一个核心挑战，得到 Koopman 算子的本征函数可以实现强非线性系统的全局线性表示。

2.3.2 受控系统下 Koopman 算子

可将 Koopman 算子推广到受控系统，对于有控制输入 u 的非线性动力学系统：

$$x_{k+1} = f(x_k, u_k) \tag{2.20}$$

其 Koopman 算子与可观测函数的关系如图2.2所示。

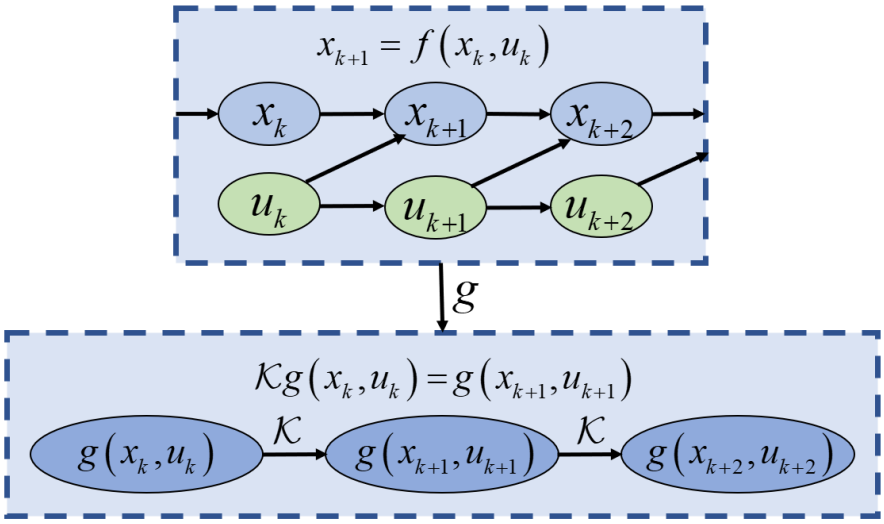

图 2.2 受控系统下 Koopman 算子与可观测函数

根据 Koopman 算子理论，对于控制输入为状态反馈形式的非线性动力学系统，通过对 Koopman 算子进行有限维近似，计算得到本征函数并合理构造可观测函数，可以将原非线性动力学系统转换为如下形式的线性系统：

$$\begin{aligned} z_{k+1} &= Az_k + Bu_k + Cw_k \\ z_k &= \Phi(x_k) \end{aligned} \tag{2.21}$$

其中 $\Phi = \begin{bmatrix} \phi_1 & \cdots & \phi_n \end{bmatrix}^{\mathrm{T}}$ 表示对系统状态进行升维的 Koopman 算子的本征函数。

第三章 基于 Q-learning 的线性二次控制

本章主要介绍了一个基于平均 Q-learning 算法的线性二次控制方法。该方法将线性二次控制理论与无模型策略迭代强化学习方法相结合，利用强化学习算法与环境交互获取带噪声的数据，结合平均成本估计值函数及状态动作值函数进行策略评估与改进，实现了噪声环境下未知模型的线性动力学系统的最优控制。随后，在动力学系统上进行了实验，测试了该算法的性能。

3.1 线性二次型最优控制问题

考虑具有高斯分布的环境噪声的线性动力学系统：

$$x_{k+1} = Ax_k + Bu_k + Cw_k \tag{3.1}$$

其中 $x_k \in \mathbb{R}^n$ 和 $u_k \in \mathbb{R}^m$ 分别代表状态和控制输入向量，n 和 m 分别表示系统的状态空间维度和控制空间维度。w_k 表示服从高斯分布 $\mathcal{N}(0, W_w)$ 的系统过程噪声，A 和 B 分别为未知的系统状态转移矩阵及控制转移矩阵。

将二次型运行成本函数定义为：

$$r(x_k, u_k) = {x_k}^{\mathrm{T}} R_x x_k + {u_k}^{\mathrm{T}} R_u u_k \tag{3.2}$$

其中 $R_x \geq 0$ 和 $R_u \geq 0$ 分别为状态加权矩阵和控制加权矩阵。当选择控制输入 u_k 为状态 x_k 的函数时，其被称为策略，本文设计形式为 $u_k = \pi(x_k) = Kx_k$ 的平稳策略来控制该动力学系统。与策略 π 相关的平均成本定义为：

$$\varepsilon(K) = \lim_{r\to\infty} \frac{1}{\tau} \mathrm{E}\left[\sum_{t=1}^{\tau} r(x_t, Kx_t)\right] \tag{3.3}$$

平均成本与系统的初始状态无关，当系统存在过程噪声时，平均成本具有非零值。进一步定义与给定策略相关的值函数：

$$V(x_k, K) = \mathrm{E}\left[\sum_{t=k}^{+\infty} \left(r(x_t, Kx_t) - \varepsilon(K)\right) | x_k\right] \tag{3.4}$$

上式中通过减去由(3.3)定义的平均成本消除过程噪声的影响，得到瞬态成本。

对于系统(3.1)，设计最优增益 K^* 使得策略 $K^* x_k$ 最小化(3.4)。任何稳定线性策略的值函数是二次的，即有

$$V(x_k, K) = x_k^{\mathrm{T}} P x_k \tag{3.5}$$

根据上述定义及 2.2 中的推导，可以得到最优策略增益 K^* 如下式：

$$K^* = -\left(R_u + B^{\mathrm{T}} P^* B\right)^{-1} B^{\mathrm{T}} P^* A$$

其中 $P^* > 0$ 是离散时间代数黎卡提方程的解，即满足

$$A^{\mathrm{T}} P^* A - P^* - A^{\mathrm{T}} P^* B\left(B^{\mathrm{T}} P^* B + R_u\right)^{-1} B^{\mathrm{T}} P^* A + R_y = 0$$

其表明设计最优增益 K^* 使得策略 $K^* x_k$ 最小化(3.4)的问题等价于经典的 LQR 问题，可以从标准的离散时间代数黎卡提方程中获得解，当状态及输入转移矩阵已知时，可由上述公式获得解决方案。而当系统模型未知时，由下文中的无模型强化学习算法迭代得到最优策略增益。

3.2 价值函数及状态-动作值函数

当动力学系统模型未知时，本文通过无模型迭代算法来解决 3.1 中描述的最优控制问题，算法需通过找到相关的价值函数 V 和状态-动作值函数 Q 来对策略进行评估并进一步进行改进。

利用(3.4)中定义的价值函数可以得到基于价值的贝尔曼方程如下：

$$V(x_k, K) = r(x_k, Kx_k) - \varepsilon(K) + \mathrm{E}\left[V(x_{k+1}, K) \,|\, x_k\right]. \tag{3.6}$$

状态动作值函数为在状态 x_k 下采取动作 u_k 然后遵循当前策略的预期成本，令 $d_k = \left[x_k^{\mathrm{T}}, u_k^{\mathrm{T}}\right]^{\mathrm{T}}$，基于状态动作值的贝尔曼方程如下式：

$$Q(x_k, u_k, K) = r(x_k, u_k) - \varepsilon(K) + \mathrm{E}\left[V(x_{k+1}, K) \,|\, d_k\right]. \tag{3.7}$$

若选择 $u_k = Kx_k$，可以从 $Q(x_k, u_k, K)$ 中得到值函数 $V(x_k, K)$。若策略增益 K 趋于稳定，则有 $Q(d_k, K) = d_k^{\mathrm{T}} G d_k$。

3.3 平均 Q-learning 算法

采取平均 Q-learning 算法对无系统模型条件下 3.1 中的最优控制问题进行求解，见算法 1，该算法为经典 Q-learning 算法的改进，通过学习相关的价值函数与状态动作值函数以及平均成本来评估策略，然后采用贪婪策略进行策略改进。该算法通过对策略评估与策略改进两个步骤进行 N 次迭代得到最优控制策略。

3.3.1 策略评估

对于第 i 轮迭代，由于稳定策略增益 K^i 有价值函数 $V^i\left(x_k, K^i\right) = x_k^{\mathrm{T}} P^i x_k$，从而可以将基于价值的贝尔曼方程(3.6)改写为如下形式：

$$x_k^{\mathrm{T}} P^i x_k = r(x_k, Kx_k) + \mathrm{E}\left[x_{k+1}^{\mathrm{T}} P^i x_{k+1} | x_k\right] - tr\left(W_w P^i\right). \tag{3.8}$$

算法 1 基于无模型强化学习的 LQR 控制策略学习

Initialize: 初始稳定策略 K^1, 初始状态 x_0, 探索协方差 W_η , 迭代次数 N, τ, τ', τ''。
for $i \in 1, ..., N$ **do**
  执行 τ 轮 $K^i x_k$ 并采集 τ 个输出样本。
  利用采集的数据构成 $\mathbf{D}^\pi$ 和 $\mathbf{D}_+^\pi$.
  利用 $\mathbf{D}^\pi$ 和 $\mathbf{D}_+^\pi$ 根据公式(3.11)、(3.12)计算 $\bar{\varepsilon}^i, \hat{P}^i$ 。
  $\mathcal{Z} = \{\}$
  for $i \in 1, ..., \tau'$ **do**
    执行 τ'' 轮 Kx 并得到最终状态 x。
    随机采样 $\eta \sim \mathcal{N}(0, W_\eta)$ 并得到 $u = Kx + \eta$。
    根据行为策略执行动作 u 并得到 x_+ 。
    将 (x, u, x_+) 添加到 $\mathcal{Z}$.
  end for
  利用数据集 $\mathcal{Z}$ 得到 $\mathbf{D}$、$\mathbf{D}_+^{\pi'}$， 根据式(3.16)计算 $\hat{G}^i$ 。
  根据贪婪策略进行策略改进 $K^{i+1} = \arg\min_u \frac{1}{i} \sum_{j=1}^{i} Q^j(x_k, u) = \sum_{j=1}^{i} -\left(\hat{G}_{22}^j\right)^{-1} \hat{G}_{12}^{jT}$.
end for

上式中 $tr\left(W_w P^i\right)$ 为遵循线性策略的平均预期成本，即 $\varepsilon^i = tr\left(W_w P^i\right)$，是关于过程噪声协方差的函数。

令 $vec(A) = \left[a_1^{\mathrm{T}}, a_2^{\mathrm{T}}, \ldots, a_l^{\mathrm{T}}\right]^{\mathrm{T}}$ 为对称矩阵 $A = [a_1, a_2, \ldots, a_l], a_i \in \mathbb{R}^p, i = 1, \ldots, l$ 的向量化形式，使得 $vec(A_1)^{\mathrm{T}} vec(A_2) = tr(A_1 A_2)$。设 $\mu(x) = vec\left(xx^{\mathrm{T}}\right)$，并且令 $\mu_k = \mu(x_k)$，$r_k = r(x_k, Kx_k)$，那么可以得到基于价值的贝尔曼方程的向量化形式为：

$$\mu_k{}^{\mathrm{T}} vec\left(P^i\right) = r(x_k, Kx_k) + \left(\mathrm{E}\left[\mu(x_{k+1}) | x_k\right] - vec(W_w)\right)^{\mathrm{T}} P^i. \tag{3.9}$$

由此可以根据策略生成数据来估计 P^i。根据策略对动力学系统执行 τ 轮 $K^i x_k$，设 $\mathbf{D}^\pi$ 为 $\tau \times n^2$ 维的矩阵，其行由向量 $\mu_1, \ldots, \mu_\tau$ 组成，类似地设 $\mathbf{D}_+^\pi$ 为 $\tau \times n^2$ 维的矩阵，其行由向量 $\mu_2, \ldots, \mu_{\tau+1}$ 组成。设 $\mathbf{W}$ 为 $\tau \times n^2$ 维的矩阵，其每一行都是向量 $vec(W_w)$，$\mathbf{r} = [r_1, \ldots, r_\tau]$。$P^i$ 的 LSTD 估计值由下式给出：

$$vec\left(\hat{P}^i\right) = \left(\mathbf{D}^{\pi\mathrm{T}}\left(\mathbf{D}^\pi - \mathbf{D}_+^\pi + \mathbf{W}\right)\right)^\dagger \mathbf{D}^{\pi\mathrm{T}} \mathbf{r}. \tag{3.10}$$

上式中 $(\cdot)^\dagger$ 表示伪逆。当噪声协方差 W_w 未知时，利用样本数据计算出经验平均成本作为 ε^i 的估计值，即

$$\bar{\varepsilon}^i = \frac{1}{\tau} \sum_{t=1}^{\tau} r_t \tag{3.11}$$

此时，P^i 的 LSTD 估计值为：

$$vec\left(\hat{P}^i\right) = \left(\mathbf{D}^{\pi\mathrm{T}}\left(\mathbf{D}^\pi - \mathbf{D}_+^\pi\right)\right)^\dagger \mathbf{D}^{\pi\mathrm{T}}\left(\mathbf{r} - \bar{\varepsilon}^i\right). \tag{3.12}$$

进一步估计状态动作值函数来评估策略。因为对于稳定稳定策略增益 K^i 有状态动作值函数 $Q^i\left(d_k,K^i\right)=d_k^{\mathrm{T}}G^i d_k$，从而可以将基于状态动作值的贝尔曼方程(3.7)改写为如下形式：

$$d_k^{\mathrm{T}}G^i d_k = r\left(x_k,u_k\right)+\mathrm{E}\left[x_{k+1}^{\mathrm{T}}P^i x_{k+1}|d_k\right]-tr\left(P^i W_w\right). \tag{3.13}$$

设 $\psi=vec\left(dd^{\mathrm{T}}\right)$，进一步可以得到基于状态动作值的贝尔曼方程的向量化形式如下：

$$\psi_k^{\mathrm{T}}vec\left(G^i\right)=r\left(x_k,u_k\right)+\left(\mathrm{E}\left(\varphi_{k+1}|d_k\right)-vec\left(W_w\right)\right)^{\mathrm{T}}vec\left(P^i\right). \tag{3.14}$$

基于上述方程，可以利用先前估计的值函数 P^i 的估计值 $\hat{P}^i$ 以及随机采样的动作来估计 G^i。在每一轮迭代中，采集 τ' 个 (x_k,u_k,x_{k+1}) 样本进行估计。在每一次采集中，先根据当前策略执行 τ'' 轮 $K^i x_k$ 得到最后的状态 x_k，然后从 $\mathcal{N}\left(0,W_\eta\right)$ 中随机采样得到 η_k，其中 W_η 是行为策略的探索协方差，执行行为策略 $u_k=K^i x_k+\eta_k$ 得到下一状态 x_{k+1}，利用采集得到的数据构成矩阵 $\mathbf{D}$ 以及 $\mathbf{D}_+^{\pi'}$。其中 $\mathbf{D}$ 是 $\tau'\times(n+m)^2$ 维矩阵，其行为向量 $\psi_1,\ldots,\psi_{\tau'}$，$\mathbf{r}=[r_1,\ldots,r_{\tau'}]$ 为其相应的成本构成的向量，$\mathbf{D}_+^{\pi'}$ 是 $\tau\times n^2$ 维的矩阵，其元素为根据行为策略采取动作后得到的下一状态向量。可以根据下式对状态动作值函数进行估计：

$$vec\left(\hat{G}^i\right)=\left(\mathbf{D}^{\mathrm{T}}\mathbf{D}\right)^{-1}\mathbf{D}^{\mathrm{T}}\left(\mathbf{r}+\left(\mathbf{D}_+^{\pi'}-\mathbf{W}\right)vec\left(\hat{P}^i\right)\right). \tag{3.15}$$

当噪声协方差 W_w 未知时，G^i 的 LSTD 估计值为：

$$vec\left(\hat{G}^i\right)=\left(\mathbf{D}^{\mathrm{T}}\mathbf{D}\right)^{-1}\mathbf{D}^{\mathrm{T}}\left(\mathbf{r}-\bar{\varepsilon}^i+\mathbf{D}_+^{\pi'}vec\left(\hat{P}^i\right)\right). \tag{3.16}$$

3.3.2 策略改进

根据估计的状态动作值函数对策略进行评估从而改进策略，令分区矩阵 $\hat{G}^i$ 为：

$$\hat{G}^i=\begin{bmatrix}\hat{G}_{11}^i & \hat{G}_{12}^i\\ \hat{G}_{12}^{iT} & \hat{G}_{22}^i\end{bmatrix} \tag{3.17}$$

其中 $\hat{G}_{11}^i\in\mathbb{R}^{n\times n},\hat{G}_{12}^i\in\mathbb{R}^{n\times m},\hat{G}_{22}^i\in\mathbb{R}^{m\times m}$ 利用贪婪策略得到改进的策略：

$$K^{i+1}=\arg\min_a\frac{1}{i}\sum_{j=1}^{i}\hat{Q}^j\left(x_k,a\right)=\sum_{j=1}^{i}-\left(\hat{G}_{22}^j\right)^{-1}\hat{G}_{12}^{jT}. \tag{3.18}$$

上式中的改进策略对先前估计值函数的平均值而不是最后一个值函数是贪婪的，这样可以避免得到不稳定的策略增益。

平均 Q-learning 算法与经典 Q-learning 算法主要有以下两个区别：一是在进行采样获取数据更新动作状态价值时，先根据目标策略对系统执行 τ'' 步控制，然后再采集第一个样本，这样使得采集到的数据都为独立样本点。二是策略增益的更新机制，经典 Q-learning 算法在更新策略时采用相对于前一个值函数贪婪的贪婪策略，而平均 Q-learning 算法中更新策略时策略相对于所有先前值函数估计值的平均值是贪婪的，这样可以使策略增益缓慢更新到最优值，当值函数估计较差时，可以避免得到不稳定的控制器增益。

3.4 实验结果

3.4.1 倒立摆控制实验

为测试算法性能，在小车倒立摆系统中进行仿真实验。由于本章为针对线性系统的基于 Q-learning 的线性二次控制算法，因此此处在小车倒立摆系统的线性仿真模型中进行测试。首先定义倒立摆系统的线性仿真模型，设重力加速度为 g，小车质量为 M，摆杆质量为 m，转动轴心到质心的距离为 l，转动惯量为 J，要控制的状态为小车的位置 s 以及摆杆与垂直方向的夹角 θ。令系统的状态向量为 $x = \begin{bmatrix} s & \dot{s} & \theta & \dot{\theta} \end{bmatrix}$，可以得到小车倒立摆系统的线性模型：

$$\dot{x} = Ax + Bu$$
$$y = Cx$$

其中

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{(m+M)mlg}{J(m+M)+Mml^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{-m^2l^2lg}{J(m+M)+Mml^2} & 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & \frac{-ml}{J(m+M)+Mml^2} & 0 & \frac{J+ml^2}{J(m+M)+Mml^2} \end{bmatrix}^{\mathrm{T}}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, D = 0.$$

利用本章中所述的无模型强化学习线性二次控制算法与上述小车倒立摆系统交互获取数据进行策略学习，令二次运行成本为：

$$r(x_k, u_k) = {x_k}^{\mathrm{T}} \begin{bmatrix} 100 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} x_k + 0.1 {u_k}^{\mathrm{T}} u_k$$

设置 $\tau = \tau' = 1000, \tau'' = 10$，迭代次数 $N = 3$。同时利用经典的基于模型的 LQR 方法进行控制，与本章中的无模型强化学习 LQR 算法进行比较，得到的控制结果如下图所示：

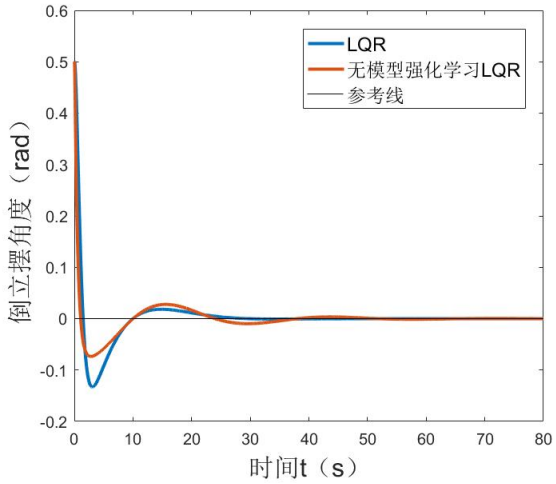

图 3.1 倒立摆角度控制

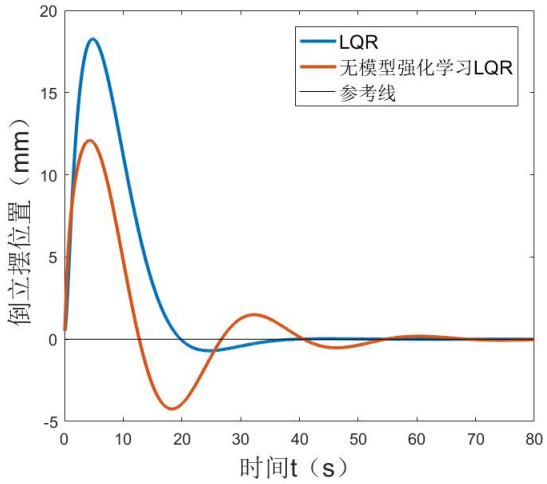

图 3.2 倒立摆位置控制

由图3.1和图3.2可以看出，本章介绍的基于改进 Q-learning 的线性二次控制算法可以成功利用与线性系统的交互数据学习最优控制策略增益，实现线性系统的二次型最优控制，并且在不利用系统模型信息的前提下达到与基于模型的经典 LQR 方法相当的控制性能。

3.4.2 干扰环境下无模型强化学习线性二次控制实验

考虑如下所示的受到干扰的线性系统：

$$\dot{x}=\begin{bmatrix}\mu & 0 & 0\\ 0 & \lambda & -\lambda\\ 2\mu & 0 & 0\end{bmatrix}x+\begin{bmatrix}1 & 0\\ 0 & 1\\ 2 & 0\end{bmatrix}u+w$$

令 $\mu=1,\lambda=-0.5$，二次运行成本设置为 $r(x_k,u_k)=x_k{}^{\mathrm{T}}x_k+0.01u_k{}^{\mathrm{T}}u_k$，$\tau=\tau'=100,\tau''=10$，迭代次数 $N=3$。利用本章介绍的无模型线性二次控制算法迭代学习最优控制策略对该系统进行反馈控制，当设置干扰 $w=0$ 即系统不受到干扰时控制效果如图3.3所示，当设置干扰不为零时，控制效果如图3.4所示：

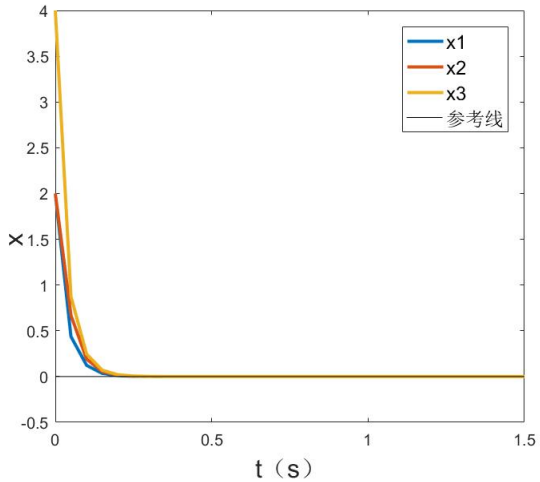

图 3.3 未受干扰控制效果

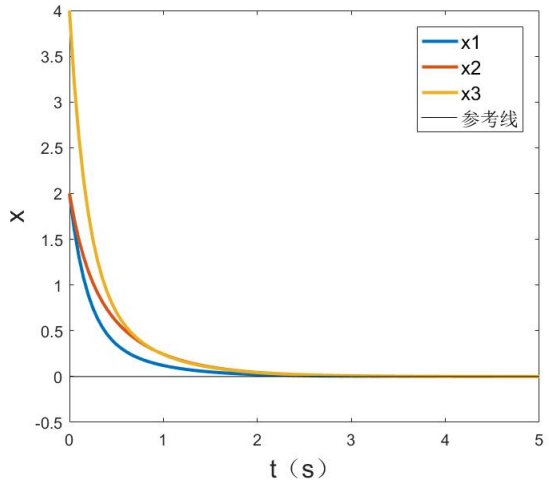

图 3.4 干扰下线性系统无模型最优控制

由图3.4可以看出在受到不确定干扰的情况下，本章介绍的基于无模型强化学习的线性二次控制算法仍然能利用带干扰噪声的交互数据学习最优控制策略增益，实现对线性系统的最优二次控制，只是与无干扰时系统的控制效果相比达到目标点的时间略有增加。

3.5 本章小结

本章主要介绍了基于无模型强化学习 Q-learning 算法的线性二次控制方法。首先对线性二次型最优控制问题的值函数等进行了定义并得到了相关的贝尔曼方程。然后介绍了一种改进的平均 Q-learning 算法，利用该无模型算法对线性二次型最优控制问题进行求解，通过策略迭代得到最优控制策略实现干扰环境下系统的最优控制。本章最后在动力学系统上测试了该算法的性能，验证了该算法可以在干扰环境下实现系统的最优控制。

第四章 基于 Koopman-无模型强化学习的非线性系统线性二次控制

本章主要提出了一种用于非线性系统最优控制的基于平均 Q-learning 算法的无模型线性二次控制方法。该方法将 Koopman 算子理论与第三章中介绍的基于 Q-learning 的线性二次控制方法相结合，利用 Koopman 可观测函数对强化学习算法与环境交互所获得的数据进行升维，从而可以将原非线性系统控制问题转换为线性二次型控制问题。然后采取无模型线性二次控制方法利用升维后的数据进行训练迭代得到最优控制策略增益，实现非线性系统的无模型强化学习线性二次控制。最后，在动力学系统上进行了实验，测试了该算法的性能。

4.1 基于 Koopman 理论的非线性系统状态升维

4.1.1 Koopman 本征函数

对于未知的非线性系统 $\dot{x} = f(x,u)$，基于 Koopman 理论进行全局线性化，合理构造 Koopman 本征函数将该非线性系统升维至线性空间，得到如下式所示的线性系统：

$$z_{k+1} = Az_k + Bu_k \tag{4.1}$$

其中 $z_k = \Phi(x_k)$，Φ 表示该机器人非线性系统的一组 Koopman 本征函数

$$\Phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_{n_{lift}} \end{bmatrix} \tag{4.2}$$

本征函数的求解视具体非线性系统而定，可以根据经验进行选择，也可以通过优化构造或神经网络等方法得到，若假设根据经验直接选择本征函数，则可以根据系统实际情况选择可观测函数为系统状态本身、状态的二次方以及径向基函数等。

4.1.2 最优化方法构造 Koopman 本征函数

给定非线性动力学系统的状态方程 $\dot{x} = f(x,u)$，在系统不受到控制时（即控制量 $u = 0$）采集数据集 $\mathcal{Z}$：

$$\mathcal{Z} = \left(\left(x_k^j \right)_{k=0}^{M_s} \right)_{j=1}^{M_t} \tag{4.3}$$

上式表示有 M_t 条不同的等距采样的轨迹数据，其中每条轨迹都有 $M_s + 1$ 个样本，即 x_k^j 上标表示轨迹，下标表示轨迹内具体时刻，x_0^j 表示第 j 条轨迹的起始点。利用该数据集可以

构造该非线性动力学系统的近似本征函数并由此得到可观测函数。

根据 Koopman 算子理论，对于连续系统有：

$$(\mathcal{K}_t\phi)(x) = e^{\lambda t}\phi(x) \tag{4.4}$$

$$(\mathcal{K}_t\phi)(x) = \phi(S_t(x)) \tag{4.5}$$

上式中 $S_t(x)$ 表示输入为 0 时动力学系统的流动，即 $\frac{d}{dt}(S_t(x)) = f(S_t(x))$，式(4.4)等价于

$$\phi(S_t(x)) = e^{\lambda t}\phi(x). \tag{4.6}$$

设 $\phi_1, ..., \phi_N$ 是 Koopman 算子的本征函数，其特征值为 $\lambda_1, ..., \lambda_N$，$\Phi = [\phi_1, \ldots, \phi_N]^{\mathrm{T}}$，设 $h(x)$ 是感兴趣的量。求解优化问题对构建的本征函数进行优化：

$$\min_{C,\Phi} \|h - C\Phi\| \tag{4.7}$$

令 $\Lambda = (\lambda_1, ..., \lambda_N)$，假设选择了边界函数 $D = (d_1, ..., d_N)$ 定义矩阵 $D(i,j) = d_i\left(x_0^j\right)$ 为数据集中轨迹起始点上的边界函数值，并且定义 $\phi_{\lambda_i,d_i}\left(x_0^j\right) := d_i\left(x_0^j\right), j = 1, \ldots, M_t, i = 1, \ldots, N.$ 那么可以得到：

$$\phi_{\lambda_i,d_i}\left(x_k^j\right) = e^{\lambda_i k T_s} D(i,j) \tag{4.8}$$

那么对本征函数的优化问题可以转化为对边界函数 d 和特征值 λ 的最优选择问题，下面介绍如何利用数据驱动算法进行边界函数的优化选择从而实现从数据中学习本征函数。令 $\mathcal{L}_\lambda$ 表示将边界函数 d 映射到本征函数 $\phi_{\lambda,d}$ 的算子，即有 $\mathcal{L}_\lambda d = e^{-\lambda\tau}(d \circ S_\tau)$，式(4.7)是非凸优化问题，但是当把 h 的各个组成部分单独考虑时，这个问题可以变成一个凸优化问题[8]。将 N 个边界函数划分为 n_h 份，$D = \left(D_1, \ldots, D_{n_h}\right)$，其中 $\sum_{i=1}^{n_h} N_i = N$，$N_i = \#D_i$ 表示 D_i 的基数。同样将特征值也划分为 n_h 份，$\Lambda = \left(\Lambda_1, \ldots, \Lambda_{n_h}\right)$，$\#\Lambda_i = N_i$。由此可以得到凸优化问题：

$$\underset{c_{i,j},d_{i,j}}{\text{minimize}} \left\| h_i - \sum_{j=1}^{N_i} c_{i,j}\mathcal{L}_{\lambda_{i,j}} d_{i,j} \right\|. \tag{4.9}$$

利用具体的数据集进行优化选择时，令 $\mathbf{h}_i$ 为一条条轨迹叠加在一起的向量，优化变量 $\mathbf{d}_{i,j}$ 是包含数据集中轨迹起点上的边界函数值的向量。令矩阵：

$$\mathbf{L}_{\lambda_{i,j}} = \mathrm{bdiag}(\underbrace{\Lambda_{i,j}, \ldots, \Lambda_{i,j}}_{\mathrm{M_t}}), \qquad \Lambda_{i,j} = \left[1, e^{\lambda_{i,j}T_s}, e^{2\lambda_{i,j}T_s}, \ldots, e^{M_s\lambda_{i,j}T_s}\right]^{\top}.$$

为简化计算过程，可将 $c_{i,j}$ 取为 1[8]，那么问题(4.9)可以转化为下式：

$$\underset{\mathbf{d}_{i,j}\in\mathbb{C}^{M_\mathrm{t}}}{\text{minimize}} \left\| \mathbf{h}_i - \sum_{j=1}^{N_i} \mathbf{L}_{\lambda_{i,j}} \mathbf{d}_{i,j} \right\|_2 \tag{4.10}$$

令

$$\mathbf{L}_{\Lambda_i}=\left[\mathbf{L}_{\lambda_{i,1}},\mathbf{L}_{\lambda_{i,2}},\ldots,\mathbf{L}_{\lambda_{i,N_i}}\right],\quad \mathbf{d}_i=\left[\mathbf{d}_{i,1}^{\top},\mathbf{d}_{i,2}^{\top},\ldots,\mathbf{d}_{i,N_i}^{\top}\right]^{\top}$$

那么问题(4.10)相当于下式：

$$\underset{\mathbf{d}_i\in\mathbb{C}^{N_iM_{\mathrm{t}}}}{\text{minimize}}\left\|\mathbf{h}_i-\mathbf{L}_{\Lambda_i}\mathbf{d}_i\right\|_2^2 \tag{4.11}$$

最小二乘问题(4.11)具有最优解:

$$\mathbf{d}_i^{\star}=\mathbf{L}_{\Lambda_i}^{\dagger}\mathbf{h}_i. \tag{4.12}$$

上述过程中采用的特征值 λ 可以通过优化来选择，式(4.11)的最小值为：

$$\|\mathbf{h}_i\|_2^2-\left\|\mathbf{L}_{\Lambda_i}\mathbf{L}_{\Lambda_i}^{\dagger}\mathbf{h}_i\right\|_2^2. \tag{4.13}$$

选择最优 $\Lambda_i=\left(\lambda_{i,1},\ldots,\lambda_{i,N_i}\right)\in\mathbb{C}^{N_i}$ 使上式最小化，该式是 Λ_i 的非凸函数，可以通过使用一组随机选择的初始条件进行局部优化来解决。经过上述过程可以成功构建非线性系统的本征函数从而可以基于由本征函数构造的可观测函数对状态向量进行升维。

4.2 平均 Q-learning 策略迭代

将非线性系统进行全局线性化后，采取基于平均 Q-learning 的无模型线性二次控制算法迭代学习最优控制策略实现最优反馈控制。根据升维后得到的线性系统，将原非线性系统控制问题转换为线性二次型控制问题，定义成本函数

$$r\left(z_k,u_k\right)={z_k}^{\mathrm{T}}R_z z_k+{u_k}^{\mathrm{T}}R_u u_k,z_k=\Phi\left(x_k\right) \tag{4.14}$$

上式中 z_k 是利用 Koopman 观测函数升维后得到的状态向量，设计形式为 $u_k=\pi\left(z_k\right)=Kz_k$ 的控制策略来控制该系统，可以得到与策略 π 相关的平均成本及值函数：

$$\varepsilon(K)=\lim_{r\to\infty}\frac{1}{\tau}\mathrm{E}\left[\sum_{t=1}^{\tau}r\left(z_t,Kz_t\right)\right] \tag{4.15}$$

$$V\left(z_k,K\right)=\mathrm{E}\left[\sum_{t=k}^{+\infty}\left(r\left(z_t,Kz_t\right)-\varepsilon(K)\right)|z_k\right] \tag{4.16}$$

利用上述定义进行无模型强化学习策略迭代算法学习最优控制策略。利用升维后的状态和输入数据找到相关的价值函数和状态-动作值函数对策略进行评估改进。根据上述定义可以得到基于价值的贝尔曼方程如下：

$$V\left(z_k,K\right)=r\left(z_k,Kz_k\right)-\varepsilon(K)+\mathrm{E}\left[V\left(z_{k+1},K\right)|z_k\right]. \tag{4.17}$$

令 $d_k=\left[z_k^{\mathrm{T}},u_k^{\mathrm{T}}\right]^{\mathrm{T}},z_k=\Phi\left(x_k\right)$，可以得到基于状态-动作值的贝尔曼方程如下：

$$Q\left(z_k,u_k,K\right)=r\left(z_k,u_k\right)-\varepsilon(K)+\mathrm{E}\left[V\left(z_{k+1},K\right)|d_k\right]. \tag{4.18}$$

首先给出初始稳定次优策略 $u_k=\pi\left(z_k\right)=K^1z_k$，然后对策略评估和策略改进步骤进行 N 次迭代，具体过程见算法 2。

算法 2 基于 koopman 理论和强化学习的非线性系统 LQR 控制策略学习

Initialize: 初始稳定策略 K^1, 初始状态 x_0, 探索协方差 W_η , 迭代次数 N, τ, τ', τ''。
构造 Koopman 可观测函数: liftfun(x_k).
for $i \in 1, ..., N$ **do**
    对当前状态进行升维 $z_k = \text{liftfun}(x_k)$.
    执行 τ 轮 $K^i z_k$ 并采集 τ 个输出样本。
    利用可观测函数将采集的样本升维并利用升维后的数据构成 $\mathbf{D}^\pi$ 和 $\mathbf{D}_+^\pi$.
    利用 $\mathbf{D}^\pi$ 和 $\mathbf{D}_+^\pi$ 根据公式(4.20)-(4.21)计算 $\bar{\varepsilon}^i$,$\hat{P}^i$ 。
    $\mathcal{Z} = \{\}$
    for $i \in 1, ..., \tau'$ **do**
        执行 τ'' 轮 Kz 并得到最终状态 x。
        将 x 升维得到 $z = \text{liftfun}(x)$.
        随机采样 $\eta \sim \mathcal{N}\left(0, W_\eta\right)$ 并得到 $u = Kz + \eta$。
        执行 u 得到 x_+ , 将 x_+ 升维得到 $z_+ = \text{liftfun}(x_+)$.
        将 (z, u, z_+) 添加到 $\mathcal{Z}$.
    end for
    利用数据集 $\mathcal{Z}$ 得到 $\mathbf{D}$、$\mathbf{D}_+^{\pi'}$， 根据 (3) 计算 $\hat{G}^i$ ， Compute $\hat{G}^i$ from $\mathcal{Z}$ using (4.23).
    根据贪婪策略进行策略改进 $K^{i+1} = \arg\min_u \frac{1}{i} \sum_{j=1}^{i} Q^j(z_k, u) = \sum_{j=1}^{i} -\left(\hat{G}_{22}^j\right)^{-1} \hat{G}_{12}^{jT}$。
end for

4.2.1 策略评估

对于第 i 轮迭代，对非线性系统执行 τ 轮 $K^i z_k$ 采集 τ 个 x_k 样本并升维得到数据进行值函数估计。令 $\mu_k = \mu(z_k), z_k = \Phi(x_k)$， $r_k = r(z_k, Kz_k)$，从而可以得到线性空间中基于价值的贝尔曼方程的向量化形式：

$$\mu_k{}^{\mathrm{T}} vec\left(P^i\right) = r(z_k, Kz_k) + \left(\mathrm{E}\left[\mu(z_{k+1}) | z_k\right] - vec(W_w)\right)^{\mathrm{T}} P^i. \tag{4.19}$$

由该贝尔曼方程可以根据数据估计值函数。在每轮迭代中对动力学系统根据当前策略执行 τ 轮 $K^i z_k$ 采集得到 τ 个 x_k 样本，利用先前得到的 Koopman 可观测函数对得到的状态样本进行升维，得到线性空间中的状态向量 $z_k = \Phi(x_k)$ 并进一步得到 μ_k。设 $\mathbf{D}^\pi$ 为 $\tau \times {n_{lift}}^2$ 维的矩阵，其中 n_{lift} 为利用 Koopman 可观测函数升维后的状态向量的维数，该矩阵的行由向量 $\mu_1, \ldots, \mu_\tau$ 组成。同样设 $\mathbf{D}_+^\pi$ 为 $\tau \times {n_{lift}}^2$ 维的矩阵，该矩阵的行由向量 $\mu_2, \ldots, \mu_{\tau+1}$ 组成，成本向量 $\mathbf{r} = [\mathbf{r_1}, \ldots, \mathbf{r}_\tau]$。

利用由升维后的样本数据得到的成本计算出经验平均成本作为平均成本的估计值：

$$\bar{\varepsilon}^i = \frac{1}{\tau} \sum_{t=1}^{\tau} r_t \tag{4.20}$$

进一步地，可以得到 P^i 的 LSTD 估计值为：

$$vec\left(\hat{P}^i\right)=\left(\mathbf{D}^{\pi\mathrm{T}}\left(\mathbf{D}^{\pi}-\mathbf{D}_{+}^{\pi}\right)\right)^{\dagger}\mathbf{D}^{\pi\mathrm{T}}\left(\mathbf{r}-\bar{\varepsilon}^i\right). \tag{4.21}$$

进一步估计状态动作值函数来评估当前策略。对于 $d_k=\left[z_k^{\mathrm{T}},u_k^{\mathrm{T}}\right]^{\mathrm{T}},z_k=\Phi\left(x_k\right)$，在稳定策略增益 K^i 下状态动作值函数满足 $Q^i\left(d_k,K^i\right)=d_k^{\mathrm{T}}G^i d_k$，令 $\psi=vec\left(dd^{\mathrm{T}}\right)$，由此可以得到基于状态-动作值的贝尔曼方程的向量化形式如下：

$$\psi_k^{\mathrm{T}}vec\left(G^i\right)=r\left(z_k,u_k\right)+\mathrm{E}\left(\varphi_{k+1}^{\mathrm{T}}|d_k\right)vec\left(P^i\right)-\varepsilon^i. \tag{4.22}$$

基于上述方程利用先前得到的值函数估计值 $\hat{P}^i$ 并通过随机采样获取动作得到数据来估计状态-动作值函数 $\hat{G}^i$。在每一轮迭代中，采集 τ' 个 (x_k,u_k,x_{k+1}) 样本并升维得到 (z_k,u_k,z_{k+1}) 进行估计。在每一轮采集中，首先根据当前策略对原动力学系统执行 τ'' 轮 $K^i z_k$ 得到最终状态 x_k，利用 Koopman 可观测函数升维得到 $z_k=\Phi\left(x_k\right)$。然后执行行为策略与原动力学系统进行交互，即从 $\mathcal{N}\left(0,W_\eta\right)$ 中随机采样得到 η_k，然后执行 $u_k=K^i z_k+\eta_k$ 得到输出的下一状态 x_{k+1}，利用 Koopman 可观测函数将该状态向量升维得到 z_{k+1}，利用升维后的数据 x_k、x_{k+1} 以及根据行为策略得到的动作 u_k 构成数据矩阵 $\mathbf{D}$ 以及 $\mathbf{D}_{+}^{\pi'}$。

设矩阵 $\mathbf{D}$ 是 $\tau'\times\left(n_{lift}+m\right)^2$ 维矩阵，其行由向量 $\psi_1,\ldots,\psi_{\tau'}$ 组成，其中 $\psi_k=vec\left(d_k{d_k}^{\mathrm{T}}\right)$，$\mathbf{r}=[r_1,\ldots,r_{\tau'}]$ 为相应成本构成的向量。设 $\mathbf{D}_{+}^{\pi'}$ 是 $\tau\times{n_{lift}}^2$ 维矩阵，其元素由根据行为策略采取动作后得到的下一状态升维后得到的状态向量构成。利用前述升维后的数据 z_k、z_{k+1} 以及根据行为策略得到的动作 u_k 构成数据矩阵 $\mathbf{D}$ 以及 $\mathbf{D}_{+}^{\pi'}$。可以通过下式得到状态-动作值函数的估计值 $\hat{G}^i$：

$$vec\left(\hat{G}^i\right)=\left(\mathbf{D}^{\mathrm{T}}\mathbf{D}\right)^{-1}\mathbf{D}^{\mathrm{T}}\left(\mathbf{r}-\bar{\varepsilon}^i+\mathbf{D}_{+}^{\pi'}vec\left(\hat{P}^i\right)\right). \tag{4.23}$$

4.2.2 策略改进

令分块矩阵 $\hat{G}^i$ 为

$$\hat{G}^i=\begin{bmatrix}\hat{G}_{11}^i & \hat{G}_{12}^i\\ \hat{G}_{12}^{iT} & \hat{G}_{22}^i\end{bmatrix} \tag{4.24}$$

其中 $\hat{G}_{11}^i\in\mathbb{R}^{n_{lift}\times n_{lift}},\hat{G}_{12}^i\in\mathbb{R}^{n_{lift}\times m},\hat{G}_{22}^i\in\mathbb{R}^{m\times m}$，$n_{lift}$ 和 m 分别为状态和控制输入的维数。利用贪婪策略得到改进的策略：

$$K^{i+1}=\arg\min_a\frac{1}{i}\sum_{j=1}^{i}\hat{Q}^j\left(z_k,a\right)=\sum_{j=1}^{i}-\left(\hat{G}_{22}^j\right)^{-1}\hat{G}_{12}^{jT}. \tag{4.25}$$

4.3 实验结果

4.3.1 慢流形非线性系统控制

为了测试基于 Koopman-无模型强化学习的非线性系统线性二次控制算法的性能，考虑非线性系统：

$$\begin{cases} \dot{x}_1 = \mu x_1 + u_1 \\ \dot{x}_2 = \lambda\left(x_2 - {x_1}^2\right) + u_2 \end{cases} \tag{4.26}$$

首先基于 Koopman 理论对该非线性系统进行升维 [37]，选择可观测函数令

$$z = \Phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ {x_1}^2 \end{bmatrix} \tag{4.27}$$

令 $\mu = 1, \lambda = -0.5$，二次运行成本设置为

$$r(z_k, u_k) = {z_k}^{\mathrm{T}} z_k + 0.01 {u_k}^{\mathrm{T}} u_k$$

$\tau = \tau' = 100, \tau'' = 10$，迭代次数 $N = 3$。利用本章介绍的无模型线性二次控制算法迭代学习最优控制策略对该系统进行反馈控制，利用学习到的策略进行轨迹跟踪的效果如图4.1和图4.2所示：

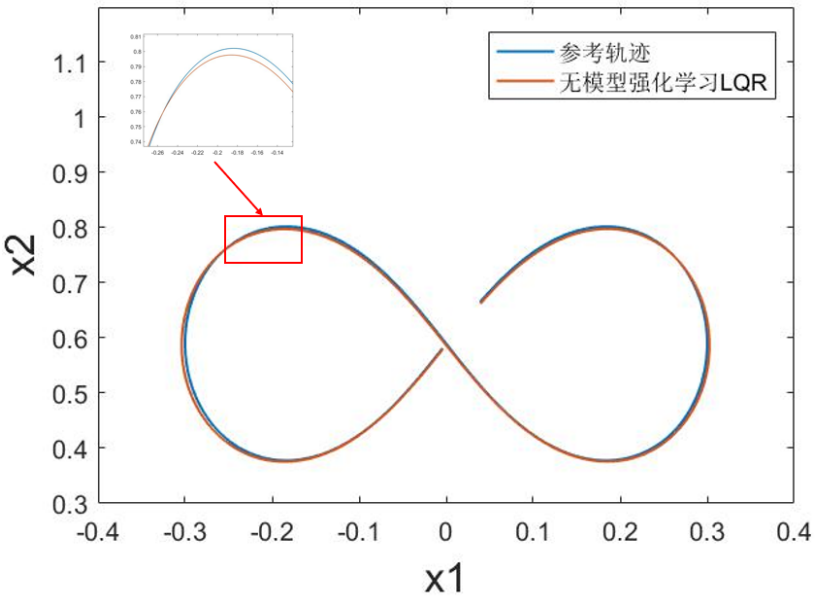

图 4.1 慢流形非线性系统线性二次控制

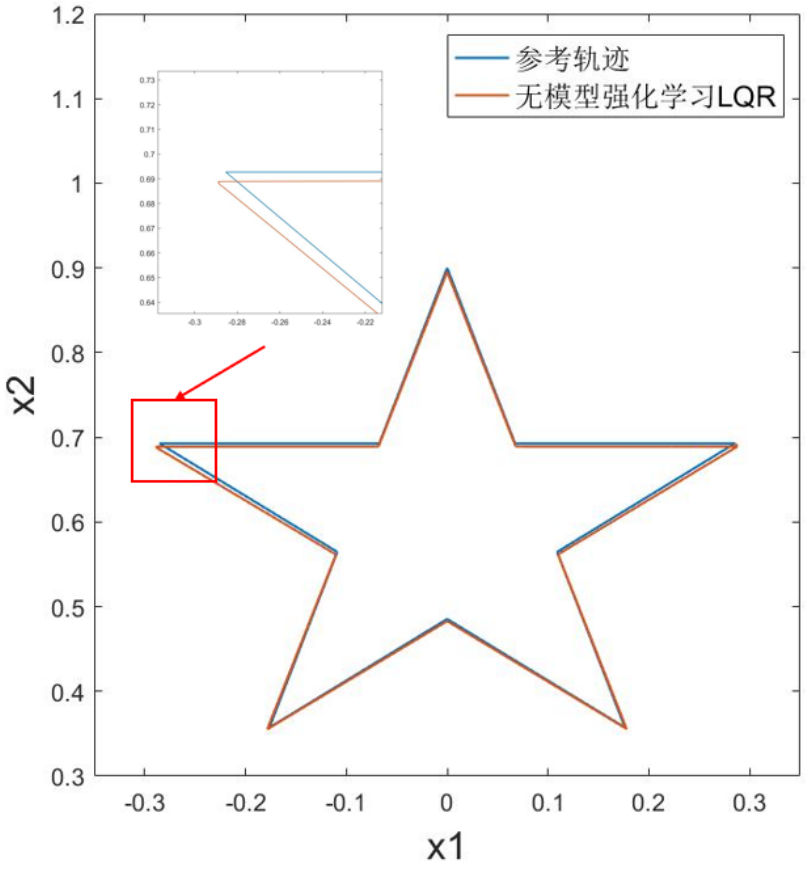

图 4.2 慢流形非线性系统线性二次控制

由图4.1和图4.2可以看到，利用本章提出的基于 Koopman 理论及无模型强化学习的线性二次控制方法成功实现了对该慢流形非线性系统的控制，在学习到的最优策略控制下状态的变化轨迹与参考轨迹几乎重合，跟踪效果良好。说明该方法可以成功通过与非线性系统交互获得的数据学习最优控制策略增益，实现对非线性系统的二次型最优控制。

4.3.2 阻尼杜芬振荡器控制

为测试干扰环境下基于 Koopman-无模型强化学习的非线性系统线性二次控制算法的性能，考虑带干扰的阻尼杜芬振荡器：

$$\begin{cases} \dot{x}_1 = x_2 + u + w_1 \\ \dot{x}_2 = -0.5x_2 + x_1 - 4x_1^3 + 0.5u + w_2 \end{cases} \tag{4.28}$$

其中 $w \in [-0.15, 0.15]$ 为有界扰动，利用最优化方法构造该非线性系统的 Koopman 本征函数，令 $n_{lift} = 10$，采集不受到控制时的数据计算得到

$$\Phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_{n_{lift}}(x) \end{bmatrix}$$

二次运行成本设置为

$$r(z_k,u_k)=z_k{}^{\mathrm{T}}R_z z_k+0.0001u_k{}^{\mathrm{T}}u_k$$

其中 R_z 为 $n_{lift}\times n_{lift}$ 维的对角矩阵，其前 $\frac{n_{lift}}{2}$ 行的对角线元素为 1，后 $\frac{n_{lift}}{2}$ 的对角线元素为 0.1。$\tau=\tau'=200,\tau''=10$，迭代次数 $N=3$。利用本章提出的基于 Koopman-无模型强化学习的线性二次控制方法学习到的策略进行轨迹跟踪，结果如图4.3所示：

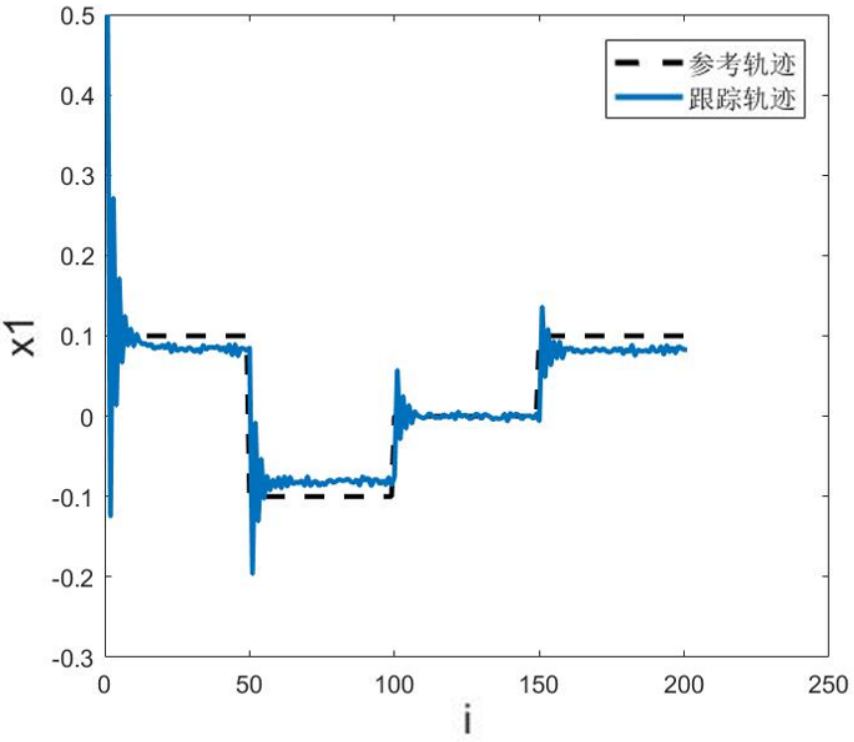

图 4.3 阻尼杜芬振荡器非线性系统无模型强化学习线性二次控制

由图4.3可以看到本章所提出的基于 Koopman-无模型强化学习的线性二次控制方法利用与干扰条件下交互得到的带噪声的数据成功学习到了阻尼杜芬振荡器的二次型最优控制策略，在学习到的策略控制下，系统状态迅速到达目标轨迹，跟踪轨迹与参考轨迹基本吻合。

4.4 本章小节

本章将 Koopman 算子理论与第三章中介绍的基于无模型强化学习的线性二次控制方法相结合，提出了基于 Koopman 理论与无模型强化学习的非线性系统线性二次控制方法。在本章中，首先对基于 Koopman 理论选取 Koopman 观测函数以及通过优化构造的方法得到 Koopman 可观测函数的过程进行了介绍。然后介绍了引入构造的可观测函数后的非线性系统最优控制策略学习过程。最后利用数值例子对所提出的方法进行了测试，通过控制慢流形非线性系统和阻尼杜芬振荡器进行轨迹跟踪测试了该方法的性能。

第五章 总结与展望

5.1 总结

非线性系统控制研究是现代控制研究理论中的重要分支，实际应用系统大多都为复杂的非线性系统，如倒立摆、机械臂等，因此对非线性系统安全有效的控制方法进行研究极具价值。现实中的复杂非线性动力学系统往往难以建立精确的数学模型，导致难以用基于模型的方法进行控制，因此本文研究基于无模型强化学习的方法学习最优控制策略。考虑干扰环境下对难以建立精确模型的非线性系统进行最优控制的问题，本文提出一种基于 Koopman 算子理论将非线性系统状态进行升维，并采用无模型强化学习算法在升维后的线性空间中学习二次型最优反馈控制策略的方法。本论文的主要工作总结如下：

1、本文研究了针对线性二次型控制问题的无模型强化学习算法，详细阐述了利用无模型强化学习算法与环境进行交互获取数据并学习状态值函数与状态-动作值函数来改进策略的方法。该方法利用根据目标策略对系统执行若干步动作后再采集数据以及更新策略增益时对先前所有值函数估计值的平均值贪婪这两个技巧对传统的 Q-learning 方法进行改进，将成本函数设置为二次型后利用改进的平均 Q-learning 算法实现了对线性系统的无模型线性二次控制。本文利用该算法在线性系统上进行了测试，结果表明该算法可以成功从数据中学习到线性系统的二次型最优控制策略。

2、本文在针对线性二次型控制问题的无模型强化学习算法的基础上引入 Koopman 算子理论，详细介绍了利用数据驱动方法最优化构造 Koopman 本征函数的过程，结合 Koopman 理论与适用于线性二次型控制问题的改进平均 Q-learning 实现对非线性动力学系统的二次型最优反馈控制。最后本文通过在慢流形非线性系统以及干扰环境下阻尼杜芬振荡器非线性系统上进行仿真实验验证了所提出方法的可行性。本文提出的基于 Koopman-无模型强化学习的非线性系统控制方法的整体研究思路如图5.1所示：

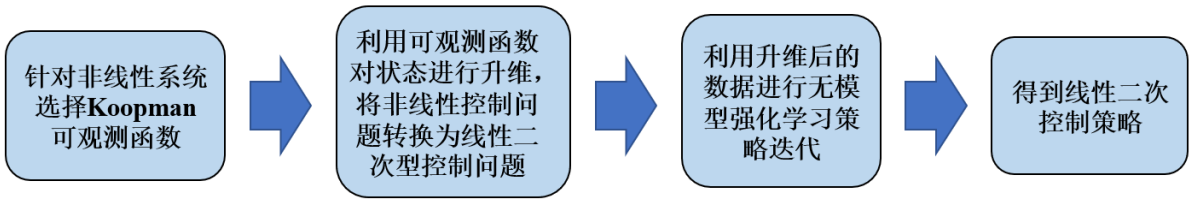

图 5.1 基于 Koopman-无模型强化学习的非线性系统线性二次控制

5.2 展望

本文所提出的方法借助无模型强化学习和 Koopman 算子理论实现了对非线性动力学系统的二次型最优控制，在此基础上可从以下几个方面展开进一步的研究：

1、本文将提出的基于无模型强化学习的非线性系统线性二次控制方法在数值例子上进行了仿真验证，但尚未应用到实际的系统中，后续工作可将该方法应用于倒立摆、机械臂等实际非线性系统的控制。

2、本文采用的无模型强化学习算法为改进的平均 Q-learning 算法，后续可以研究使用其他无模型强化学习算法或训练技巧进行改进，对比分析各方法的控制效果。

参考文献

[1] Shangtai J, Zhongsheng H, Weihong W. Model-free adaptive control used in permanent magnet linear motor[C]. 2007. 748–751.

[2] Yaghmaie F A, Gustafsson F, Ljung L. Linear quadratic control using model-free reinforcement learning[J]. *IEEE Transactions on Automatic Control*, 2023, 68(2):737–752.

[3] Abbasi-Yadkori Y, Lazic N, Szepesvari C. Model-free linear quadratic control via reduction to expert prediction[J]. 2018.

[4] Matni N, Proutiere A, Rantzer A, et al. From self-tuning regulators to reinforcement learning and back again[C]. 2019. 3724–3740.

[5] Bian T, Jiang Y, Jiang Z P. Adaptive dynamic programming for stochastic systems with state and control dependent noise[J]. *IEEE Transactions on Automatic Control*, 2016, 61(12):4170–4175.

[6] Bahare, Kiumarsi, Frank, et al. H control of linear discrete-time systems: Off-policy reinforcement learning - sciencedirect[J]. *Automatica*, 2017, 78:144–152.

[7] Mauroy A, Mezic I, Susuki Y. The koopman operator in systems and control concepts, methodologies, and applications: Concepts, methodologies, and applications[J]. *Lecture Notes in Control and Information Sciences*, 2020.

[8] Korda M, Mezi I. Optimal construction of koopman eigenfunctions for prediction and control[J]. *IEEE Transactions on Automatic Control*, 2020, 65(12):5114–5129.

[9] Iqbal J, Ullah M, Khan S G, et al. Nonlinear control systems –a brief overview of historical and recent advances[J]. *Nonlinear Engineering*, 2017, 6.

[10] Hamza M F, Yap H J, Choudhury I A, et al. Current development on using rotary inverted pendulum as a benchmark for testing linear and nonlinear control algorithms[J]. *Mechanical Systems and Signal Processing*, 2019, 116:347–369.

[11] 范昕. 基于自适应动态规划的一类非线性系统控制 [D]. 南京邮电大学, 2022.

[12] Werbos P J. Advanced forecasting methods for global crisis warning and models of intelligence[J]. *general systems yearbook*, 1977.

[13] Wen G, Chen C L P, Ge S S, et al. Optimized adaptive nonlinear tracking control using actor–critic reinforcement learning strategy[J]. *IEEE Transactions on Industrial Informatics*, 2019, 15(9):4969–4977.

[14] 闫珍珍. 基于神经网络的受扰非线性系统控制 [D]. 西安电子科技大学, 2021.

[15] Ouyang Y, Dong L, Sun C. Critic learning-based control for robotic manipulators with prescribed constraints[J]. *IEEE Transactions on Cybernetics*, 2022, 52(4):2274–2283.

[16] Sutton R, Barto A. Reinforcement learning:an introduction[M]. Reinforcement Learning:An Introduction, 1998.

[17] 孙悦雯, 柳文章, 孙长银. 基于因果建模的强化学习控制: 现状及展望 [J]. 自动化学报, 2023, 49(661-677).

[18] Watkins C. Learning with delayed rewards[J]. *Ph.D. thesis, Cambridge University*, 1989.

[19] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. *Computer Science*, 2013.

[20] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. 2017.

[21] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. *Computer ence*, 2015.

[22] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[Z], 2018.

[23] Gu S, Holly E, Lillicrap T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]. 2017. 3389–3396.

[24] 杨永亮. 无模型自适应动态规划及其在多智能体协同控制中的应用 [D]. 北京科技大学, 2018.

[25] Xi A, Mudiyanselage T W, Tao D, et al. Balance control of a biped robot on a rotating platform based on efficient reinforcement learning[J]. *IEEE/CAA Journal of Automatica Sinica*, 2019, 6(4):938–951.

[26] 许雅筑，武辉，游科友，宋士吉. 强化学习方法在自主水下机器人控制任务中的应用 [J]. 中国科学*:* 信息科学, 2020, 50(1798-1816).

[27] Koopman B O. Hamiltonian systems and transformation in hilbert space.[J]. *Proceedings of the National Academy of Sciences*, 1931, 17(5):315–318.

[28] Budii M, Mohr R, Mezi I. Applied Koopmanisma[J]. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2012, 22(4). 047510.

[29] Rowley C W, Mezi? I, Bagheri S, et al. Spectral analysis of nonlinear flows[J]. *Journal of Fluid Mechanics*, 2009, 641:115.

[30] Williams M O, Kevrekidis I G, Rowley C W. A data-driven approximation of the koopman operator: Extending dynamic mode decomposition[J]. 2014.

[31] Bethany, Lusch, Nathan J, et al. Deep learning for universal linear embeddings of nonlinear dynamics.[J]. *Nature Communications*, 2018.

[32] 王琦，杨毅远，江季. Easy rl：强化学习教程 [M]. 北京: 人民邮电出版社, 2022.

[33] Norvig S R . Artificial intelligence: A modern approach[M]. 北京: 清华大学出版社, 2006.

[34] 张伟楠，沈键，俞勇作. 动手学强化学习 [M]. 北京: 人民邮电出版社, 2022.05.

[35] 于程隆. 基于模型学习和线性二次型最优控制的机械臂控制器设计 [D]. 哈尔滨工业大学, 2018.

[36] Korda M, Mezi I. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control[J]. *Automatica*, 2018, 93:149–160.

[37] Kutz S L B J N. Data-driven science and engineering: Machine learning, dynamical systems, and control[M]. Cambridge University Press, 2019.02.

致　　谢

夏天又快要到了，和四年前的那个夏天一样炽热。那时的我憧憬着即将到来的大学生活，欣喜又忐忑，转眼间步入大学的尾声，回首过往有收获也有遗憾。我始终觉得自己是幸运的，在一路长大的过程中遇到了很多很好的人，收获了满满的善意与美好，我很庆幸在我生命的不同时刻，恰好有你们在。

感谢阎石老师和赵东东老师，在我对未来感到迷茫而不知所措的时候接纳了我。两位老师给了我许多的指导与帮助，每次组会上耐心地指出我的问题，在我不自信的时候给予我鼓励和建议。经师易遇，人师难求，两位老师不仅在专业上给予了指导，还在交流时告诉我们人生的道理，令我受益匪浅。真诚地感谢两位老师的包容与指导。感谢杨肖迪学长，在我进入本研计划以来给了我许多鼓励与帮助，不厌其烦地回答我的问题。愿万事胜意!

感谢永远坚定地支持我的家人，始终温暖着我让我有勇气向前。感谢我最最爱的外婆陪伴我一路成长，用她不自知的浪漫和温柔温暖着我，让我无论何时想起都会感到治愈。感谢我的妈妈和爸爸，总是鼓励我支持我，尊重我做的所有决定，告诉我开心最重要。感谢悉心照顾我教导我的小姨，我始终记得在我低落难过时小姨和小姨父的那句“我们也永远是你的后盾”，万分感激，不胜言表。

感谢我的挚友李琴、独孤、蒋猪、龙龙，大学四年我们分隔几地，很庆幸没有走散，很庆幸有你们总是积极地回应我关于点滴小事的分享，在我情绪失控的时候打着电话陪我，包容我的坏脾气。感谢陪我一起走过这四年的李艳、柴姐、吕琪，因为你们我的大学时光才变得丰富而美好。朋友是自己选的家人，我真的非常非常希望我们能像家人一样互相陪伴着一直走下去。

感谢大学最后时刻的一场相遇，所有美好的相遇都为时未晚。

最后感谢努力向前，热爱生活的自己，“我永远都没有长大，但我永远都没有停止过成长。”

论文（设计）成绩

导师评语

冯敏同学基于 Koopman 算子理论和无模型强化学习对非线性系统控制展开了研究。论文详细分析了现有方法的局限性，针对实际复杂非线性系统难以建立精确的数学模型等问题，提出了基于无模型强化学习的非线性系统线性二次控制方法。相关实验结果表明，提出的无模型方法达到了控制要求。

论文结构合理，层次分明，叙述准确，图表规范，实验设计合理，工作量饱满，完成了开题报告的内容，达到了本科生毕业论文水平。

建议成绩 优 指导教师（签字） 阎石

答辩小组意见

经答辩小组一致讨论，该论文通过答辩，成绩为 优

答辩委员会负责人（签字）

成绩 优秀 学院（盖章）

2023 年 6 月 27 日